bio-protocol Preprint

Not Peer Reviewed
Ask a question online

TCGA and Sanger database analysis

MP Miquel Angel Pujana MB Mary Helen Barcellos-Hoff

Updated date: Nov 6, 2021

An abbreviated version of this protocol was published in Science Translational Medicine in Feb 2021

Loss of TGFβ signaling increases alternative end-joining DNA repair that sensitizes to genotoxic therapies across cancer types

DOI: 10.1126/scitranslmed.abc4465

Detailed protocol

GenScore Analytical Protocol

This protocol has been implemented using Open Source R programming language.

Step 1: Data download and expression matrices

Gene expression data from the TCGA GBM, LUSC and OV cancer datasets was downloaded from the following sources:

- GBM: Downloaded from "UCSC Xena cancer browser" using R package "UCSCXenaTools". Contains expression measured by Microarra AffyU133a. Unit: log2(affy RMA). Dimensions: 539 samples x 12042 genes.
- LUSC: Downloaded from "UCSC Xena cancer browser" using R package "UCSCXenaTools" in August 2020. Contains expression measured b IlluminaHiSeq_RNASeqV2. Unit: pan-cancer normalized log2(norm_count+1). Dimensions: 553 samples x 20530 genes.
- OV: Downloaded from GDC using the R package "TCGAbiolinks" in January 2020. Contains expression measured by Agilent Microarray (AgilentG4502A_07_3). Dimensions: 562 samples x 16210 genes.

All TCGA cancers (including the above) data were also downloaded from GDC Data Portal. FPKM-UQ RNASeq files of selected samples were added to the cart and downloaded directly or (if large files) using the GDC Data Transfer tool.
Each dataset consisted of a gene expression matrix containing samples as rows, identified by their TCGA ID, and genes as columns, identified by their name.
Gene expression values were scaled into gene-centered z-scores, by first transposing the gene expression matrix and then applying the R function *scale*.

Step 2: GenScore value of signatures

The *gsva* function of the *GSVA* package (version 1.34) is used. This package is available and freely downloadable from the Bioconductor library.
The *genScore* Genetic Score metric library is used. This can be downloaded and installed from GitHub (https://github.com/pujana-lab/genScore). The library has been developed *ad hoc* for this project.
Process:
1. ssGSEA values:

a. The expression data and gene signatures are loaded separately.
   I. The expression data is defined in matrix format, where the rows correspond to genes and columns to samples. This matrix is saved as variable named *maseq_matrix*.
   II. The gene signatures are defined in list format with elements corresponding to signature's genes. The signatures are saved as variable *signatures list*.
b. The *ssgsea* method is executed. Their first two parameters are the gene expression matrix (*maseq_matrix*) and gene signatures (*signatures_list*). This method calculates a gene set enrichment score per sample as the normalized difference in empirical cumulative distribution functions (CDFs) of gene expression ranks inside and outside the gene set, following the steps described in (Barbie et al. 2009).
c. The output of this procedure is a matrix with two rows and as many columns as samples.

2. genScore values:
The example provided in the *genScore* website may be followed to compute the values. The method does not require any specific parameter (*genScore::genScore(ssgsea$up_tgfb, ssgsea$alt_ej*), for example).

If wished to group samples by tertiles, the *categorizeSamples* method from the same library can be used. This method requires three variables: the array of the signature, the lower threshold and the upper threshold. Thus, to extract the lower and upper tertiles the following command can be used: *genScore::categorizeSamples(scores, lowThreshold=1/3, highThreshold=2/3)* High βAlt and LowβAlt: this classification omits the samples that are in the middle group.

Step 3: Survivals

A survival data file is needed. In the case of TCGA studies, we have used TCGA Pan-Cancer Clinical Data Resource published at Liu et al '18 (https://pubmed.ncbi.nlm.nih.gov/29625055/). Alternatively, cBioPortal data was used: LUSC (lusc_tcga_pan_can_atlas_2018_clinical_data), GBM (gbm_tcga_pan_can_atlas_2018_clinical_data), OV: (ov_tcga_pan_can_atlas_2018_clinical_data).

The algorithms of *Surv*, *survfit,* and *coxph* in the *survival* package are used, as well as the *ggsurvplot* function from the *survminer* package to draw the survival curves. Both packages are available from CRAN (https://cran.r-project.org/).
The patient/tumor groups obtained in the previous step are used in an automated way to compare the High βAlt with the Low βAlt survivals.
Cox survival models use Low βAlt as a reference group and include, whenever possible, the covariates age, and tumor grade/stage. The *ggsurvplot* algorithm is applied to compute log-rank tests.

Related files

TCGA-signatures_protocol_110621.zip

How to cite:(Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Pujana, M. and Barcellos-Hoff, M. (2021). TCGA and Sanger database analysis. Bio-protocol Preprint. bio-protocol.org/prep1428.
2. Liu, Q., Palomero, L., Moore, J., Guix, I., Espín, R., Aytés, A., Mao, J., Paulovich, A. G., Whiteaker, J. R., Ivey, R. G., Iliakis, G., Luo, D., Chalmers, A. J., Murnane, J., Pujana, M. A. and Barcellos-Hoff, M. H.(2021). Loss of TGFβ signaling increases alternative end-joining DNA repair that sensitizes to genotoxic therapies across cancer types . Science Translational Medicine 13(580). DOI: 10.1126/scitranslmed.abc4465